CFO 12088 US
08/863,047

日 本 国 特 許 庁
PATENT OFFICE
JAPANESE GOVERNMENT

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。
This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日
Date of Application: １９９７年 ９月 ５日

出 願 番 号
Application Number: 平成 ８年特許願第２３２９６９号

出 願 人
Applicant (s): キヤノン株式会社

１９９７年 ６月２０日

特 許 庁 長 官
Commissioner,
Patent Office 荒井寿光

出証番号 出証特平０９－３０４６８０８

【書類名】　特許願

【整理番号】　3328019

【提出日】　平成 8年 9月 3日

【あて先】　特許庁長官　荒井　寿光　殿

【国際特許分類】　G06F 15/20

【発明の名称】　情報検索方法及びその装置、記憶媒体

【請求項の数】　7

【発明者】

【住所又は居所】　東京都大田区下丸子３丁目３０番２号キヤノン株式会社内

【氏名】　伊藤　史朗

【発明者】

【住所又は居所】　東京都大田区下丸子３丁目３０番２号キヤノン株式会社内

【氏名】　大谷　紀子

【発明者】

【住所又は居所】　東京都大田区下丸子３丁目３０番２号キヤノン株式会社内

【氏名】　柴田　昇吾

【発明者】

【住所又は居所】　東京都大田区下丸子３丁目３０番２号キヤノン株式会社内

【氏名】　上田　隆也

【発明者】

【住所又は居所】　東京都大田区下丸子３丁目３０番２号キヤノン株式会社内

【氏名】　池田　裕治

【特許出願人】

【識別番号】　000001007

【郵便番号】　146
【住所又は居所】　東京都大田区下丸子３丁目３０番２号
【氏名又は名称】　キヤノン株式会社
【代表者】　御手洗　冨士夫
【電話番号】　03-3758-2111
【代理人】
【識別番号】　100069877
【郵便番号】　146
【住所又は居所】　東京都大田区下丸子３丁目３０番２号キヤノン株式会社内
【弁理士】
【氏名又は名称】　丸島　儀一
【電話番号】　03-3758-2111
【手数料の表示】
【予納台帳番号】　011224
【納付金額】　21,000円
【提出物件の目録】
【物件名】　明細書　1
【物件名】　図面　1
【物件名】　要約書　1
【包括委任状番号】　9003707
【プルーフの要否】　要

【書類名】　　明細書

【発明の名称】　情報検索方法及びその装置、記憶媒体

【特許請求の範囲】

【請求項１】　情報を集合として保持する情報保持手段と、

検索条件を保持する検索条件保持手段と、

前記情報保持手段に保持されている各要素から、前記検索条件保持手段に保持されている検索条件を満足する要素を検索する検索手段と、

前記検索条件の検索結果を用いて、前記情報保持手段に保持されている各情報集合のスコアを演算する演算手段と、

前記演算手段により演算されたスコアを保持する集合スコア保持手段と、

を有することを特徴とする情報検索装置。

【請求項２】　前記演算手段は、集合の要素数と集合中の検索条件を満足する要素数とから集合全体のスコアの統計的区間推定を行いその下限値をもってスコアとすることを特徴とする請求項１に記載の情報検索装置。

【請求項３】　前記演算手段は、集合の要素数と各要素の検索条件に対するスコアを用いて集合全体のスコアの統計的区間推定を行いその下限値をもってスコアとすることを特徴とする請求項１に記載の情報検索装置。

【請求項４】　情報を集合として保持する情報保持工程と、

検索条件を保持する検索条件保持工程と、

前記情報保持工程により保持されている各要素から、前記検索条件保持工程により保持されている検索条件を満足する要素を検索する検索工程と、

前記検索工程の検索結果を用いて、前記情報保持工程により保持されている各情報集合のスコアを演算する演算工程と、

前記演算工程により演算されたスコアを保持する集合スコア保持工程と、

を有することを特徴とする情報検索方法。

【請求項５】　前記演算工程は、集合の要素数と集合中の検索条件を満足する要素数とから集合全体のスコアの統計的区間推定を行いその下限値をもってスコアとすることを特徴とする請求項４に記載の情報検索方法。

【請求項６】　前記演算工程は、集合の要素数と各要素の検索条件に対するスコアを用いて集合全体のスコアの統計的区間推定を行いその下限値をもってスコアとすることを特徴とする請求項１に記載の情報検索方法。

【請求項７】　情報を集合として保持する情報保持工程と、

検索条件を保持する検索条件保持工程と、

前記情報保持工程により保持されている各要素から、前記検索条件保持工程により保持されている検索条件を満足する要素を検索する検索工程と、

前記検索工程の検索結果を用いて、前記情報保持工程により保持されている各情報集合のスコアを演算する演算工程と、

前記演算工程により演算されたスコアを保持する集合スコア保持工程と、

を有することを特徴とする情報検索プログラムを格納したコンピュータで読み取り可能な記憶媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、情報の集合を検索する情報検索方法及びその装置に関するものである。

【０００２】

【従来の技術】

情報を含有する文書やレコードなどを集合として保持し、情報を利用する場合には集合として検索する装置がある。例えば、従来の文書処理装置（特願平８－１２９８９９）では、文書を集合として保持するものとしてフォルダを用意し、フォルダ単位で検索を行なっている。

【０００３】

従来、この種の文書処理装置では、検索条件に対する集合のスコアとして検索条件を満足する文書やレコードが集合中に含まれる割合を採用していた。この割合が高いほど、集合として検索条件に合う可能性が高いと考えられる。

【０００４】

【発明が解決しようとしている課題】

しかしながら、上記従来例の装置では、集合中の要素数の多寡がスコアに反映されないため、要素数が異なる集合同士の比較がしづらいという問題点がある。例えば、集合の要素数が１で条件を満足する要素が１の場合と、集合の要素数が１０で条件を満足する要素が９の場合では、前者のスコアが高くなるが、集合として検索条件に合う可能性が高いのは後者である。

【０００５】

本発明は上記の問題に鑑みてなされたものであり、集合の要素数の多寡を反映して、集合のスコアを算出する情報検索方法及び装置を提供することを目的とする。

【０００６】

【課題を解決する為の手段】

上記課題を解決する為に、本発明は、情報を集合として保持する情報保持手段と、検索条件を保持する検索条件保持手段と、前記情報保持手段に保持されている各要素から、前記検索条件保持手段に保持されている検索条件を満足する要素を検索する検索手段と、前記検索手段の検索結果を用いて、前記情報保持手段に保持されている各情報集合のスコアを演算する演算手段と、前記演算手段により演算されたスコアを保持する集合スコア保持手段とを有することを特徴とする情報検索装置を提供する。

【０００７】

【発明の実施の形態】

以下、図面を参照して本発明の実施形態を詳細に説明する。

【０００８】

図１は、本発明の一実施形態に係る装置の基本構成を示すブロック図である。

【０００９】

同図において、１０１は検索対象文書を保持する文書保持部である。１０２は文書の集合を保持する文書集合保持部である。１０３は検索条件を保持する検索条件保持部である。１０４は文書保持部１０１に保持されている文書から、検索

条件保持部１０３に保持されている検索条件を満足する文書を検索する文書検索部である。１０５は文書検索部１０４の検索結果を保持する検索結果保持部である。１０６は検索結果保持部１０５に保持されている検索結果を用いて、文書集合保持部１０２に保持されている各文書集合のスコアを計算する集合スコア演算部である。１０７は集合スコア演算部１０６により計算されたスコアを保持する集合スコア保持部である。

【００１０】

図２は、本実施形態の情報検索装置の具体的構成を示す図である。

【００１１】

同図において、２０１はＣＰＵであり、後述する手順を実現するプログラムに従って動作する。２０２はＲＡＭであり、検索条件保持部１０３と検索結果保持部１０５と集合スコア保持部１０７と上記プログラムの動作に必要な記憶領域とを提供する。２０３はＲＯＭであり、後述する手順を実現するプログラムを保持する。２０４はディスク装置であり、文書保持部１０１と文書集合保持部１０２を実現する。２０５はバスである。２０６は後述するプログラムにより検索された検索結果を表示するＣＲＴである。２０７は検索条件等を入力するキーボード（ＫＢ）である。

【００１２】

なお、本実施例では、文書集合保持部１０２は、文書集合に一意の集合番号を付与して、集合番号が示す文書集合に属する文書の文書番号のリストを保持する構造になっている。図４に文書集合保持部の例を示す。列４０１には集合番号が保持され、列４０２には文書番号のリストが保持されている。

【００１３】

この他、文書保持部１０１は、文書に一意の文書番号を付与して、文書番号が示す文書本文を保持する構造になっている。検索条件保持部１０３は、検索語のリストを保持する構造になっている。検索結果保持部１０５は、文書番号のリストを保持する構造になっている。集合スコア保持部１０７は、集合番号が示す文書集合のスコアを保持する構造になっている。

【0014】

以下、図3のフローチャートを参照して、本実施形態の情報検索装置における処理の手順を示す。

【0015】

ステップS301では、検索条件保持部103に検索語のリストから成る検索条件cが保持されているか否かを調べ、保持されている場合はステップS302に移る。保持されていない場合は、ステップS301を繰り返す。

【0016】

ステップS302では、文書保持部101中の文書から、検索条件保持部103に保持されている検索条件cを満足する文書を検索する。文書の本文中にc中の各語が出現するか否かをパターンマッチングにより調べて、全ての語が出現する場合に、その文書は検索条件cを満足すると判断する。検索条件を満足する文書の文書番号を検索結果保持部105に保持する。そして、ステップS303に移る。

【0017】

ステップS303では、kの値を1に設定する。そして、ステップS304に移る。

【0018】

ステップS304では、kの値と文書集合保持部102中の集合数Nを比較し、$k \leq N$であればステップS305に移る。$k > N$であれば処理を終了する。

【0019】

ステップS305では、文書集合保持部102中のk番目の文書集合D_kのスコアs_kを計算する。まず、文書集合D_kに属する文書数をnとし、D_kに属する文書のうち、検索結果保持部105中にも含まれる文書の数をxとする。その上で、自由度（ϕ_1，ϕ_2）のF分布を用いて、次のようにs_kを計算する。

【0020】

【外1】

$$s_k = \frac{\phi_2}{\phi_1 F_{\phi_1}^{\phi_2}(\alpha)+\phi_2}$$

ここで、自由度は、$\phi_1 = 2(n-x+1)$，$\phi_2 = 2x$である。

【００２１】

また、αは区間推定における信頼度を指定するパラメータであるが、例えば０．１とする。そして、ステップＳ３０６に移る。

【００２２】

ステップＳ３０６では、ステップＳ３０５で求めたスコアs_kを集合スコア保持部１０７に保持する。そして、ステップＳ３０４に戻る。

【００２３】

例えば、図４の文書集合保持部の例において、ステップＳ３０２が終了した後、検索結果が（１，３，５）となったとすると、文書集合ごとのｎとｘの値は、D_1ではｎ＝５，ｘ＝３、D_2ではｎ＝１，ｘ＝１、D_3ではｎ＝３，ｘ＝１となる。従って、各文書集合のスコアs_kは、

【００２４】

【外２】

$$s_1 = \frac{6}{6F_6^6(0.1)+6} \simeq 0.25$$

$$s_2 = \frac{2}{2F_2^2(0.1)+2} \simeq 0.10$$

$$s_3 = \frac{2}{6F_6^2(0.1)+2} \simeq 0.03$$

となる。

【００２５】

このように本実施形態の情報検索装置を用いると、検索条件により合う文書集合に対して高いスコアが与えられるので、得られたスコアを用いると、利用者が検索条件に合う文書集合を探すことが容易になる。

【００２６】

（第２の実施形態）

上記実施形態においては、集合の要素数と集合中で検索条件を満足する要素数とから２項分布の統計的区間推定を行ないその下限値をもって集合全体のスコア

とする場合について説明した。

【００２７】

本実施形態では、集合の要素数と各要素の検索条件に対するスコアを用いて母平均の区間推定を行ないその下限値をもって集合全体のスコアとする場合について説明する。

【００２８】

なお、本実施形態の基本構成は図１と同様である。ただし、文書検索部１０４では、各文書の検索条件に対するスコアを算出し、検索結果保持部１０５では、文書ごとのスコアを保持する。図６に検索結果保持部１０５の例を示す。列６０１は文書番号を保持し、列６０２がその文書のスコアを保持する。

【００２９】

以下、図５のフローチャートを参照して、本実施形態の情報検索装置における処理の手順を示す。

【００３０】

ステップＳ５０１では、検索条件保持部１０３に検索語のリストから成る検索条件ｃが保持されているか否かを調べ、保持されている場合はステップＳ５０２に移る。保持されていない場合は、ステップＳ５０１を繰り返す。

【００３１】

ステップＳ５０２では、文書保持部１０１中の文書について、検索条件保持部１０３に保持されている検索条件ｃに対するスコアを計算する。スコアは、文書の本文中にｃ中の各語が出現する頻度を用いて一般的に計算される。計算されたスコアを検索結果保持部１０５に保持する。そして、ステップＳ５０３に移る。

【００３２】

ステップＳ５０３では、ｋの値を１に設定する。そして、ステップＳ５０４に移る。

【００３３】

ステップＳ５０４では、ｋの値と文書集合保持部１０２中の集合数Ｎを比較し、$k \leq N$であればステップＳ５０５に移る。$k > N$であれば処理を終了する。

【００３４】

ステップＳ５０５では、文書集合保持部１０２中のｋ番目の集合D_kに属する文書数をｎとする。また、D_kに属する文書のスコアの平均をｘとする。ｎ＞１の場合は、不偏推定量Ｖを次の式により求め、

【００３５】

【外３】

$$V = \frac{\Sigma(x-\bar{x})^2}{n-1}$$

その上で、自由度ϕ、両側確率αのｔ分布を用いて、次のようにs_kを計算する。

【００３６】

【外４】

$$s_k = \bar{x} - t(\phi, \alpha)\frac{\sqrt{V}}{\sqrt{n}}$$

ここで、自由度は、ϕ＝ｎ－１である。ｎ＝１の場合は、

【００３７】

【外５】

$$s_k = \alpha\,\bar{x}$$

とする。なお、αは区間推定における信頼度を指定するパラメータであるが、例えば０．１とする。そして、ステップＳ５０６に移る。

【００３８】

ステップＳ５０６では、ステップＳ５０５で求めたスコアｓを集合スコア保持部１０７に保持する。そして、ステップＳ５０４に戻る。

【００３９】

上記実施形態においては、検索条件として検索語のＡＮＤ条件をとる場合について説明したが、これに限定されるものではない。他の論理関係や文書中の位置条件など、文書に対する任意の検索条件を用いてよい。

【００４０】

また、パターンマッチングにより文書の検索を行なう場合について説明したが、これに限定されるものではない。文書に対するインデックスを作成しておき、インデックスを用いて検索するなど、任意の検索手段を用いてよい。

【００４１】

また、集合を構成する情報が文書である場合について説明したが、これに限定されるものではない。データの集合であるレコードなど任意のものでよい。この場合、情報に応じた検索手段を用いる。

【００４２】

また、各集合についてスコアを計算する場合について説明したが、これに限定されるものではない。文書ごとに、その文書が属する集合を保持しておき、検索結果に含まれる文書を最低１個含む集合についてだけスコアを計算してもよい。それ以外の集合のスコアは０になる。

【００４３】

また、集合に対して全てのスコアを保持する場合について説明したが、これに限定されるものではなく、一部のスコアだけを保持してもよい。例えば、スコアが予め設定された閾値を越えたものだけを保持する方法や予め設定された数や割合の範囲内で結果を保持する方法がある。

【００４４】

また、各部を同一の計算機上で構成する場合について説明したが、これに限定されるものではなく、ネットワーク上に分散した計算機や処理装置などに分かれて各部を構成してもよい。

【００４５】

また、図２は、本実施形態の情報検索装置の具体的構成を示す図である。

【００４６】

検索条件保持部と検索結果保持部と集合スコア保持部をＲＡＭで、文書保持部と文書集合保持部をディスク装置で実現する場合について説明したが、これに限定されるものではなく、任意の記憶媒体を用いて実現してもよい。

【００４７】

また、プログラムをＲＯＭに保持する場合について説明したが、これに限定されるものではなく、任意の記憶媒体を用いて実現してもよい。また、同様の動作をする回路で実現してもよい。

【００４８】

また、本発明は、前述した実施形態の機能を実現するソフトウェアのプログラムコードを記録した記憶媒体を、システムあるいは装置に供給し、そのシステムあるいは装置のコンピュータ（またはＣＰＵやＭＰＵ）が記憶媒体に格納されたプログラムコードを読出し実行することによっても、達成されることは言うまでもない。

【００４９】

この場合、記憶媒体から読出されたプログラムコード自体が本発明の新規な機能を実現することになり、そのプログラムコードを記憶した記憶媒体は本発明を構成することになる。

【００５０】

プログラムコードを供給するための記憶媒体としては、例えば、フロッピーディスク、ハードディスク、光ディスク、光磁気ディスク、ＣＤ－ＲＯＭ、ＣＤ－Ｒ、磁気テープ不揮発性のメモリカード、ＲＯＭなどを用いることができる。

【００５１】

また、コンピュータが読出したプログラムコードを実行することによって、前述した実施形態の機能が実現される他、そのプログラムコードの指示に基づき、コンピュータ上で稼働しているＯＳなどが実際の処理の一部または全部を行い、その処理によっても前述した実施形態の機能が実現され得る。

【００５２】

さらに、記憶媒体から読出されたプログラムコードが、コンピュータに挿入された機能拡張ボードやコンピュータに接続された機能拡張ユニットに備わるメモリに書込まれた後、そのプログラムコードの指示に基づき、その機能拡張ボードや機能拡張ユニットに備わるＣＰＵなどが実際の処理の一部または全部を行い、その処理によっても前述した実施形態の機能が実現され得る。

【００５３】

尚、本発明は、前述した実施形態の機能を実現するソフトウェアのプログラムコードを記録した記憶媒体からそのプログラムをパソコン通信等通信ラインを介して要求者にそのプログラムを配信する場合にも適用できることは言うまでもない。

【００５４】

【発明の効果】

以上説明したように、本発明によれば、集合として保持されている情報に対して集合の要素数の多寡を反映して集合のスコアを算出するようにしたので、検索条件により合う文書集合に対して高いスコアが与えられ、利用者が検索条件に合う文書集合を探すことが容易になるという効果が得られる。

【図面の簡単な説明】

【図１】

本発明に係る情報検索装置の実施形態の基本構成を示すブロック図である。

【図２】

本発明の実施形態の具体的構成を示す図である。

【図３】

本発明の実施形態における処理の概要を示すフローチャートである。

【図４】

本発明の実施形態における文書集合保持部の例を示す図である。

【図５】

本発明の第２の実施形態における処理の概要を示すフローチャートである。

【図６】

本発明の第２の実施形態における検索結果保持部の例を示す図である。

【図７】

本発明の制御プログラムをコンピュータにロードする様子を示す図である。

【符号の説明】

２０１　ＣＰＵ

２０２　ＲＡＭ

２０３　ＲＯＭ

２０４　ディスク装置

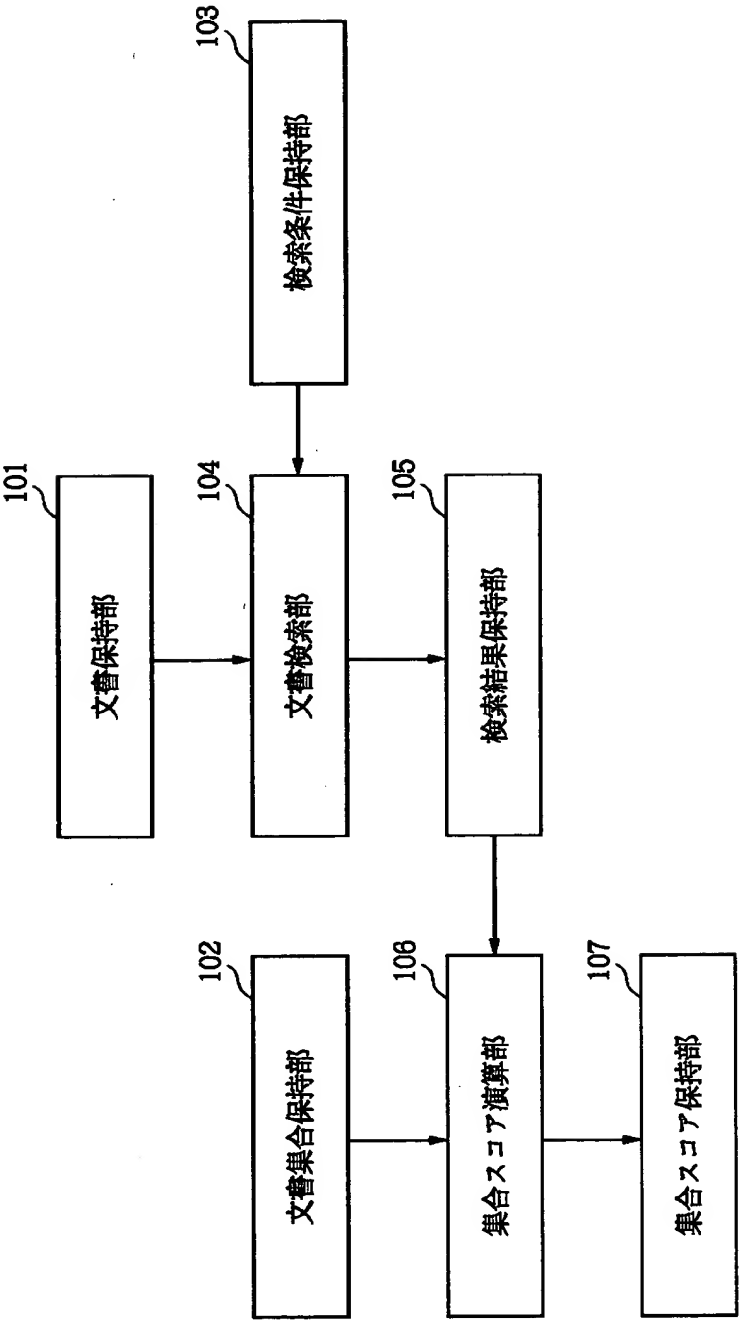
２０５　バス

２０６　ＣＲＴ

２０７　ＫＢ

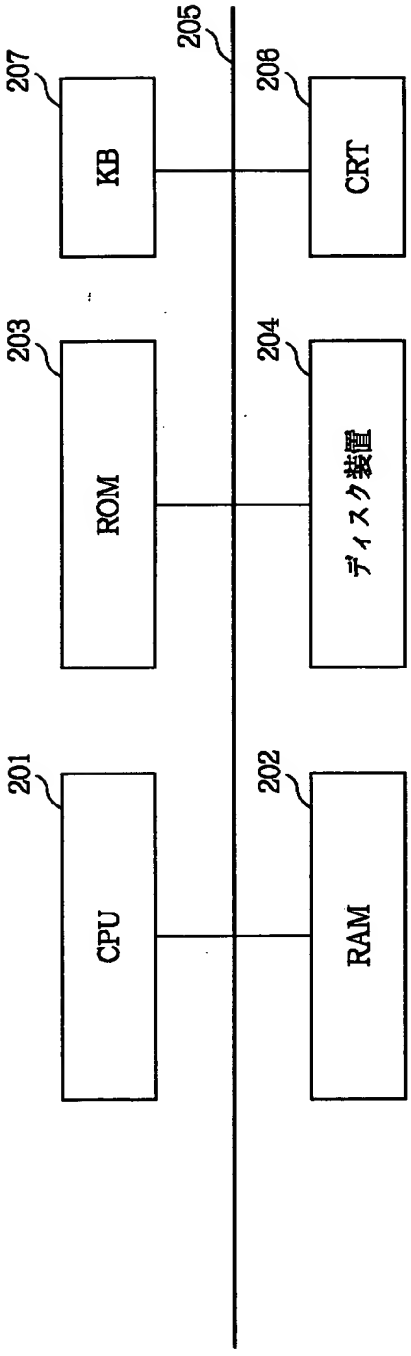
【書類名】　　図面

【図１】

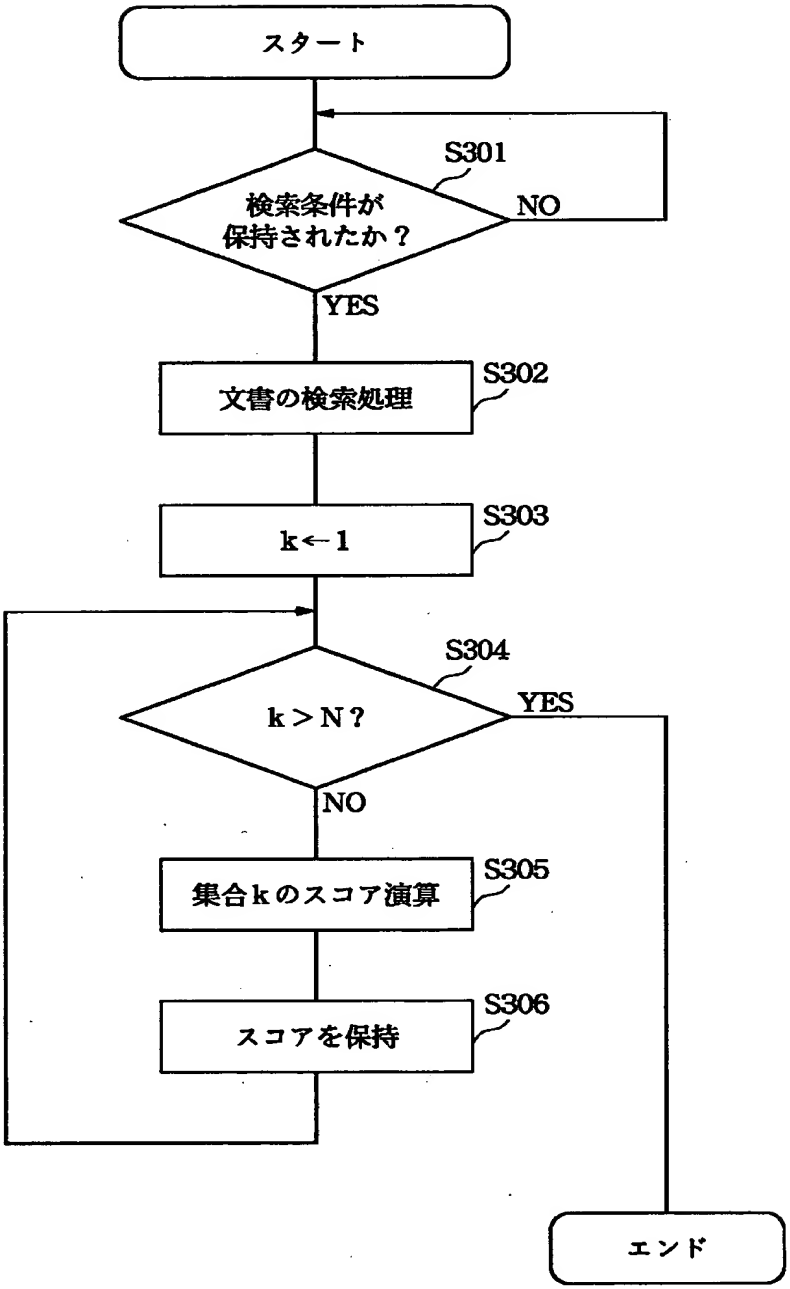


出証特平０９－３０４６８０８

【図2】



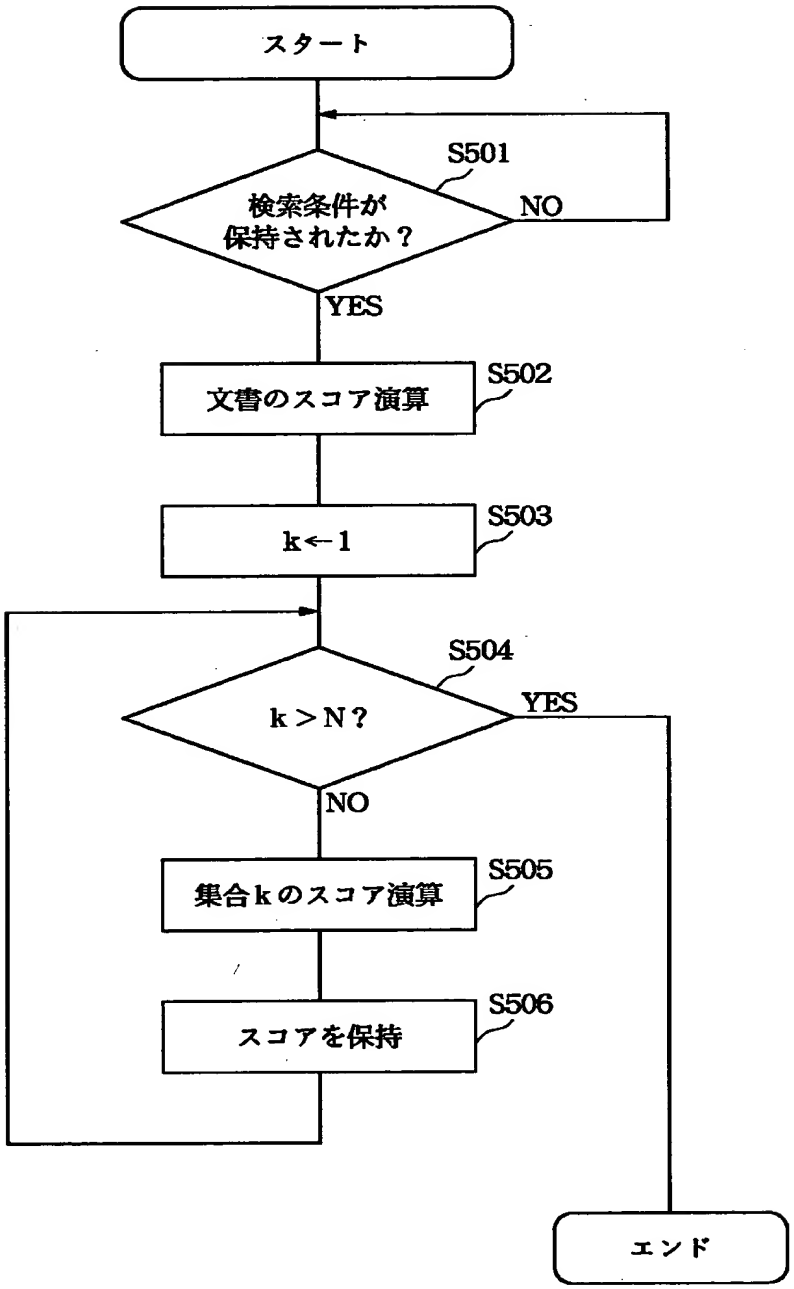
【図３】


スタート
S301
検索条件が
保持されたか？
NO
YES
文書の検索処理
S302
k←1
S303
S304
k＞N？
YES
NO
集合kのスコア演算
S305
スコアを保持
S306
エンド

【図４】

401	402
1	1, 2, 3, 4, 5
2	3
3	2, 5, 6

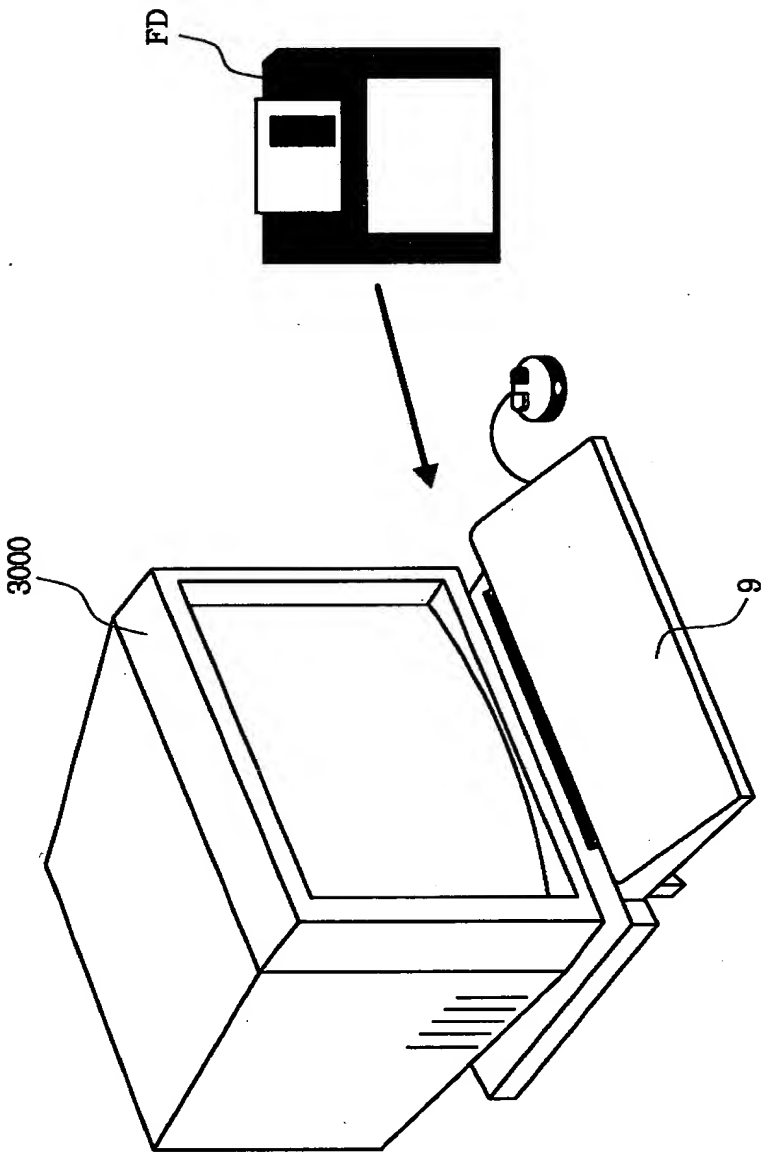
【図５】


スタート
S501
検索条件が
保持されたか？
NO
YES
文書のスコア演算
S502
k←1
S503
S504
k＞N？
YES
NO
集合kのスコア演算
S505
スコアを保持
S506
エンド

【図6】

601	602
1	0.8
2	0.1
3	0.95
4	0.0
5	0.85
6	0.2

【図７】



【書類名】　　要約書

【要約】

【課題】　文書等の情報を集合とした集合単位で検索してスコアを求める場合、要素数の多寡によってはスコアが適切とならないことがある。

【解決手段】　文書保持部１０１の文書から、検索条件保持部１０３に保持されている検索条件を満足する文書を検索する。検索条件を満足する文書の文書番号を検索結果保持部１０５に保持する。文書集合毎の文書数と検索結果保持部１０５中に含まれる文書の数を用いて、集合スコア演算部１０６により、各文書集合スコアを求めることにより、検索条件により合う文書集合に対して高いスコアが与えられる。

【選択図】　　図１

【書類名】　職権訂正データ
【訂正書類】　特許願

＜認定情報・付加情報＞
【特許出願人】
【識別番号】　000001007
【住所又は居所】　東京都大田区下丸子３丁目３０番２号
【氏名又は名称】　キヤノン株式会社
【代理人】　申請人
【識別番号】　100069877
【住所又は居所】　東京都大田区下丸子３－３０－２　キヤノン株式会社内
【氏名又は名称】　丸島　儀一

出　願　人　履　歴　情　報

識別番号　　〔０００００１００７〕

１．変更年月日　　１９９０年　８月３０日
［変更理由］　新規登録
住　所　　東京都大田区下丸子３丁目３０番２号
氏　名　　キヤノン株式会社